基于关联规则的文本主题深度挖掘应用研究*

阮光册　夏　磊
(华东师范大学信息管理系　上海 200241)
(上海图书馆　上海 200031)

摘要: 【**目的**】准确理解文本信息中潜在的知识关联，丰富文本知识挖掘的方法。【**方法**】将主题模型和关联规则相结合，运用 LDA 主题模型抽取文本中的主题集合，在实现文本降维的同时，实现文本在语义空间的表达；通过关联规则进一步挖掘文本中主题的语义关联。【**结果**】设置合理的支持度和置信度阈值，可以有效地挖掘文本中潜在知识的关联，实现对文本的深入“理解”。【**局限**】数据预处理过程中，用户自定义词典的设计会对实验结果产生一定的影响。【**结论**】提出一种非结构化文本信息潜在语义关联挖掘的新思路，改善了针对文本信息知识发现的效果。

关键词: 关联规则　主题模型　文本主题
分类号: G350

1 引　言

随着信息技术和互联网通信技术的发展与普及，产生了大量的文本信息，文本信息的快速增长使得人们在信息处理和检索中面临前所未有的挑战。对文本的理解，不仅有助于信息检索、内容发现等情报工作的开展，同时对信息的有效分类、组织也提供了借鉴。然而文本信息的组织形式是松散的，对一般用户来说，过量的文本信息反而使得信息使用率降低，使人们迷失在复杂的信息空间中[1]。海量的文本信息已经远远超出人们的理解和概括能力，通过人工的方式去查找有用的信息并凝练知识已变得不可能，如何利用计算机有效地组织和管理这些文本资源，并运用信息技术帮助用户在大量文本中挖掘隐含的知识成为当前信息技术领域面临的一大挑战。

随着对文本认识的发展，人们开始追求对文本本身更深的理解，从而使计算机甚至人们能够更好地“理解”文本。对文本的深入理解一方面可以完成文本挖掘或自然语言处理，并实现如自动人工问答等信息服务。另一方面也能挖掘文本潜在语义，为信息工作者提供相关的技术支持。在主题模型出现以前，信息处理和文本挖掘领域对文本的表示主要采用空间向量模型[2]和统计语言模型[3]。这两种方式虽然在方法上存在差异，但也有很多相同点，都能够将一个文档实现“文本→词”的映射或表示。

传统的文本表示方法将文本表现在词典空间上，这种方式会忽略文本中很多重要的信息，无法达到文本语义的理解。主题模型引入语义维度，将文本信息在语义层实现了浓缩，即实现“文本→语义→词”的语义映射。本文将关联规则与主题模型相结合，从文本的主题模型入手，构建大量文本的主题集合，通过关联规则算法，构建文本主题的关联关系，实现对文本主题的深度挖掘。并以有关“一带一路”的新闻报道文本为例，实现文本信息的主题关联挖掘实验。

通讯作者：阮光册，ORCID：0000-0001-8685-5234，E-mail：rgc1976@126.com。
*本文系上海哲学社会科学一般项目“基于主题模型的学科交叉知识发现研究”(项目编号: 2016BTQ002)的研究成果之一。

2 相关研究

关联规则最初应用于购物篮问题分析[4]，通过交易数据库中频繁购买模式挖掘不同商品间的关联关系，发现隐藏在数据中的有价值知识。随着关联规则应用的深入研究，各种改进和扩展关联规则的算法被应用于诸多领域数据集的频繁模式挖掘中，用以揭示事物间隐含的关联。在改进算法的同时，关联规则也被应用到文本分析领域[5]，主要有两种方法：基于关键字的文本关联规则挖掘；借助领域本体进行文本关联规则挖掘。

(1) 基于关键字的文本关联规则挖掘通常分为两个步骤：挖掘文本集中频繁共现的关键词，形成频繁项集；发现关键词频繁项集间的关联规则。文献[6]将文本集中的文本作为事务，文本中的词作为项，将句子作为文本基本语义单元，借助句子中单词的共现关系，寻找最大关联的关键词组，生成关联规则。文献[7]采用关联规则挖掘分子生物领域的文本，通过识别文本中的关键词，寻找关联规则。文献[8]则利用文本集中词的共现程度寻找关联规则。文献[9]利用关联规则挖掘中文文本的主题词，通过构建候选关键词的二元组，过滤掉根本不可能成为关键词汇的词性组合。然而，由于高频关键词存在孤立性，因此在发现文本语义知识层的规则时存在一定的不足。

(2) 在本体领域的文本关联规则挖掘，主要是通过构建领域本体，对文本进行概念抽取，寻找概念关系组合，统计后打分，找出各层次间概念的关联规则。文献[10]通过构建“hotel”领域本体，采用人工和机器相结合的方法，半自动化抽取文本中信息，进而挖掘频繁的概念组合，得到文本之间的层次关联关系。文献[11]构建足球评论的领域本体，通过分析评论文本的语言特征，挖掘文本中动名词三元组，得到概念频繁集的组合，生成足球评论的关联规则。基于领域本体的文本关联挖掘的特点是将关键词抽象到概念高度寻找关联，但是本体需要领域专家建立，应用文本类型少，影响了挖掘方法的通用性。

综上所述，目前关联规则在文本挖掘中的应用主要还是对文本关键词或概念进行挖掘，缺乏对文本语义层次的理解。挖掘大规模文本集合中隐含的知识关联仍然存在一定的困难。

为此，本文从文本的主题模型入手，将文本的词项空间变换为主题空间，实现对文本的语义降维，再进行关联规则挖掘，通过控制关联规则算法的支持度和置信度，挖掘文本主题的关联关系，得到更深层次的文本知识。

3 研究基础

3.1 主题模型

主题模型[12]在自然语言处理领域备受关注，主题可以看成是词项的概率分布，通过词项在文档级的共现信息抽取出语义相关的主题集合，得到文本在低维空间中的表达。主题(Topic)被看作是文本包含词项的概率分布，主题模型假设一篇文档中的单词可以交换次序而不影响模型的训练结果，这个假设即词袋(Bag of Words)。

通常文本中出现的词汇都可以表达其主题，只不过与主题的相关程度有所不同。LDA 模型是一个包括了单词层、主题层、文档层的三层贝叶斯概率模型。假设在一个文档集 D 中有 m 篇文档，即 $D=\{d_1, d_2, d_3, \cdots, d_m\}$，文档集 D 中分布着 k 个主题 Z，即$\{Z_1, Z_2, Z_3, \cdots, Z_k\}$，其中每个主题 Z 都是一个基于单词集合$\{w_1, w_2, \cdots, w_n\}$的概率多项分布，W 则是所有描述主题的单词构成的词汇集合。

3.2 关联规则

关联规则旨在从大量数据中发现事务之间有趣的关联关系，以揭示隐藏其中的行为模式。关联规则挖掘通过用户指定最小支持度和最小置信度寻找事务的某些关联关系。关联规则的处理可以分成两个步骤：识别频繁项目集；挖掘关联规则。

假设关联规则挖掘的事务集合记为 D，$D=\{t_1, t_2, \cdots, t_k, \cdots, t_n\}$，则 t_k $(k=1, 2, \cdots, n)$称为事务(Transactions)。而 $t_k=\{i_1, i_2, \cdots, i_m, \cdots, i_p\}$，$i_m$ $(m=1, 2, \cdots, p)$则称为项目(Item)。设 $I=\{i_1, i_2, \cdots, i_m\}$是 D 中全体项目组成的集合，I 的任何子集 X 称为 D 中的项目集(Itemset)。

对于任意目标集 X、Y，若 $X\subset I$，$Y\subset I$，并且 $X\cap Y=\phi$，X、Y 之间的关联性用 $X\Rightarrow Y$ 表示，则：

$$\mathrm{support}(X\Rightarrow Y)=\mathrm{count}(X\Rightarrow Y)/|D| \quad (1)$$

$$\mathrm{confidence}(X\Rightarrow Y)=\mathrm{support}(X\Rightarrow Y)/\mathrm{support}(X) \quad (2)$$

其中，支持度 $\mathrm{support}(X\Rightarrow Y)$表示 X、Y 共同出

现的比例，用来衡量 $X \Rightarrow Y$ 在事务集 D 中的显著性；置信度 confidence($X \Rightarrow Y$)表示在条件概率 $P(Y \mid X)$下，用来衡量 $X \Rightarrow Y$ 在目标事务集中的显著性[13]。

3.3 基于关联规则的文本主题深度挖掘

文本信息往往是围绕某个主题展开，针对某一领域，信息之间往往存在着直接或间接的语义关联，识别文本中具有语义关联的实体，将有助于人们对文本集的认识，并能够更好地理解文本集中隐含的知识。

文本的主题模型实现了文本在语义空间上的表述，基于关联规则的文本主题深度挖掘则希望对文本所包含的主题进行关联规则发现，计算文本中实体间语义关联的强度，将关联强度大的主题进行描述。

假设文本空间 D 上有主题集合 T 和词汇集合 W，其中 $D=\{d_1, d_2, \cdots, d_i\}$，$d_i$ 代表第 i 篇文本，$T=\{t_1, t_2, \cdots, t_k\}$，$t_k$ 代表文本空间中的第 k 个主题，$w_{ik} \in W$ 为第 i 篇文本所包含的主题词项。在关联规则处理中，将 D 表示为交易组，d_i 为交易组中的第 i 项交易，由唯一的交易标识(TID)和一组项列表(Itemlist)组成，W 则为项目集，由描述 D 集合的主题词项组成，包含 w_{ik} 的交易集合表示为 $\{d_i|W \subseteq d_i, d_i \in D\}$。

为此，基于关联规则的文本主题深度挖掘的基本思路如图 1 所示：

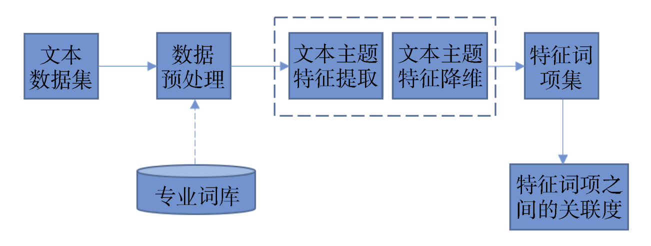


图 1 基于关联规则的文本主题挖掘

由图 1 可见，首先获取所需的文本数据集。针对文本数据集的特征，构建相应的专业词库，便于进行分词、停用词等预处理。预处理后的文本集通过主题模型对其进行主题提取，然后根据主题模型生成的“文本–主题”分布特征文件，选取高概率主题的特征词对文本进行描述，进而实现文本基于主题特征的降维。文本主题特征降维的目的不仅减小了文本表示模型的特征向量的维数，同时保留了文本的语义特征，从而提高信息提取的效率和精度。降维后的文本形成了特征词项集合，这里可以将该集合看作关联规则的交易组，集合中的每条文本则是交易组中的交易，采用关联规则算法，通过设置合理的支持度和置信度阈值，能实现文本集合中主题关联的识别，进而实现文本集合潜在知识的发现。

将关联规则和主题模型结合，对文本进行主题深度挖掘的优势主要体现在：

(1) 解决了关键词之间的语义关系。传统的关键词提取大多为使用统计方法提取文本中的术语，然而高频术语有可能是一个单纯的词项，与文本中其他词项之间缺乏语义联系。

(2) 实现文本在语义空间的降维描述。LDA 主题模型作为一种降维工具，在主题求解过程中，通过机器学习，能够得到一个文档在主题空间的表示。此过程将词项空间的文档转换成主题空间的表示，有效地实现了文本维度的降低。

(3) 发现词项之间的知识关联。关联规则算法通过支持度和置信度的设置，能够实现多元词汇关联的挖掘，实现信息之间直接或间接的关联发现。

4 实验与讨论

4.1 实验数据及实验步骤

在国家图书馆慧科报刊数据库中，以题名和主题检索包含“一带一路”的相关文献，获取 2014 年全年有关“一带一路”政策的新闻报道，下载量为 13 392 篇，共计 73.7MB。

在数据集预处理过程中，本文构建了自定义词典、去停用词等，并通过 Python 和 Jieba 分词组件对文本集进行分词处理。对于自定义用户词典，通过人工分析，提炼文本集中的专业词汇，如“一带一路”、“海上丝绸之路”、“丝绸之路经济带”、“中国梦”等词项生成自定义词典，在预处理中，这些词将不做分词处理；为有效降低文本的维数，分词时定义停用词表，去除如：“记者”、“日报”、“晚报”、“本报讯”等新闻报道中出现的高频无意义词汇。

4.2 文本主题挖掘

文本主题的挖掘是实验的基础，主题识别的效果将影响关联规则的实现。首先将总文本量的 1/3 作为主题进行学习和训练，然后对剩余文本进行主题识别，并获得相应的文本降维描述。

文本的主题是文本内容的抽象描述，在 LDA 模型中，主题 T 的数量需要预先给定，通常语料集越大主题的数量越多。本文使用统计语言模型中常用的评价指标即困惑度(Perplexity) [12]确定最优的主题数。困惑度是文档集中包含的各句子相似性几何均值的倒数，随句子相似性的增加而逐步递减。困惑度表示预测数据时的不确定度，取值越小表示性能越好。图 2 显示了对文本集合困惑度计算的结果。迭代 1 000 次，每个主题选择 10 个主题词。

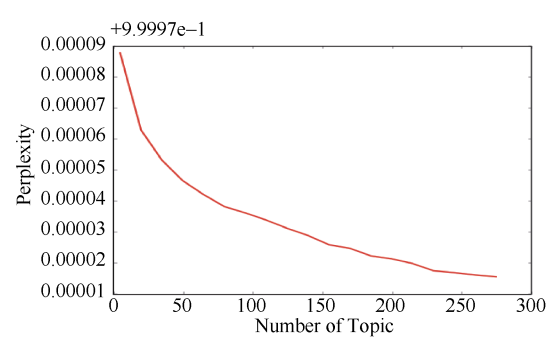


图 2　困惑度计算结果

从图 2 可以看出，困惑度曲线在 230 的位置有一个转折点，随后趋于平稳，因此本文在主题模型计算中选择主题数量为 230 个。其他参数为，∂ =50/230，β =0.01，k=230，每个主题选 10 个主题词，迭代 1 000 次。

主题模型求解后，获取每篇高概率的主题作为文本的降维描述。对每篇文本所计算的主题中，选取概率最高的前 3 个主题作为对该篇文本的表示。降维后，每篇文本将由 30 个主题词项在语义维度上进行描述。实验的部分结果如图 3 所示:

1　方式　模式　全球　未来　影响　环境　带来　市场　阶段
作用　合作　丝绸之路经济带　海上丝绸之路　提出　经济
丝绸之路　构想　国家　建设　贸易 两国　领域　关系　合作
中方　建设　一带一路　中国　发展　主席
2　基础设施　基金　国家　中国　一带一路　建设　项目　投资
成立　丝路 合作　丝绸之路经济带　海上丝绸之路　提出　经济
丝绸之路　构想　国家　建设　贸易　有望　公司　证券
行业　股份　受益　板块　基建　上市公司　投资
3　工作　推进　精神　发展　一带一路　建设　领导　加快
改革　中央 方式　模式　全球　未来　影响　环境　带来
市场　阶段　作用　合作　丝绸之路经济带　海上丝绸之路　提出
经济　丝绸之路　构想　国家　建设　贸易
4　经济带　长江　区域　规划　一带一路　建设　国家　发展
地区　推进 工作　推进　精神　发展　一带一路　建设
领导　加快　改革　中央 合作　丝绸之路经济带　海上丝绸之路
提出　经济　丝绸之路　构想　国家　建设　贸易
5　物流　港口　运输　口岸　货物　铁路　港　集装箱
通道　航线 产业　打造　重点　创新　项目　城市　基地
加快　生态　提升　历史　文化　世界　传承　起点　文化遗产
丝路　城市　古代　艺术

图 3　文本降维后的描述(部分)

4.3　基于关联规则的主题深度挖掘

对文本主题关联的挖掘，本文采用 Apriori 算法。根据上述文本降维的结果，每篇文本作为一项事务 t_k，其中 $t_k=\{w_1, w_2, \cdots, w_i\}$，$w_i$描述的是文本中第 i 个主题词项，对应关联规则中的一个项目。

采用 R 语言对降维后的文本主题数据进行关联规则分析，待分析数据的基本信息如表 1 所示:

表 1　待挖掘数据的基本信息

项目	说明
文本数量(Row)	13391 行
主题词数量(Item)	1469 项
稀疏矩阵(Sparse Matrix)	380717
密度(Density)	0.01935385
平均每篇文本的主题词数量	28.43

运用关联规则进行关联挖掘过程中，单纯设定最小支持度和最小置信度可能会产生一些价值并不大的规则。为了有效解决这个问题，文献[14]引入改善度(lift)的概念。改善度是采用相关分析描述规则内在价值的度量，并描述项集 X 对 Y 的影响力的大小。项集{X}和项集{Y}之间的改善度可表示为如下公式:

$$\mathrm{lift}(X \to Y) = \frac{\mathrm{support}(X \cup Y)}{\mathrm{support}(X) \times \mathrm{support}(Y)} \qquad (3)$$

可知，当 lift(X→Y)=1，表明{X}、{Y}相互独立，说明两个事件没有任何关联；如果该值小于 1，则表明两个事件之间是互相排斥的；一般认为，当 lift 的值大于 3 时，挖掘的关联规则是有价值的。

为此，本文在实验确定文本关联规则支持度的时候，集合了改善度 lift 的概念。通过实验，不同支持度和置信度的分布如图 4 所示。

图 4(a)和图 4(b)的支持度值分别为 0.1 和 0.2，置信度值为 80%；图 4(c)和图 4(d)的支持度值分别为 0.1 和 0.2，置信度的值为 95%。可以看出，当支持度设定为 0.1 时，图 4(a)和图 4(c)产生的关联规则(rules)均超过 10 万条，支持度设定为 0.2 时，两个实验(图 4(b)和图 4(d))产生的规则也均超过 1 万条。从规则可视化分布来看，图 4(b)绝大多数的高强度关联规则的 lift 值超过 3，且具有更高的置信度。因此本文在文本主题关联挖掘中采用图 4(b)的参数设定进行。

通过实验，本文共得到 10 228 条规则，平均置信度为 0.9982，平均改善度为 3.862。在对规则进行排序处理，对高关联规则的信息进行人工处理后，得到的信息如表 2 所示。

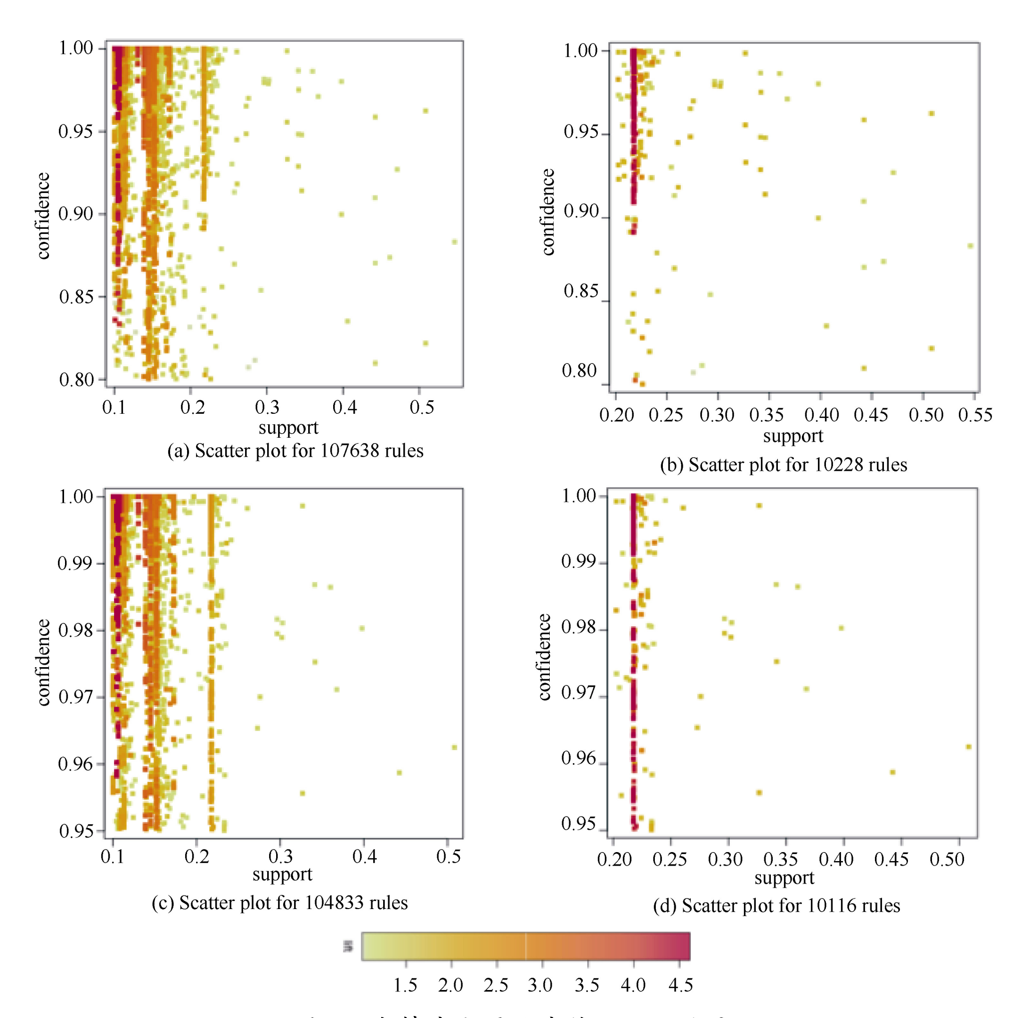


图 4 支持度和置信度值可视化效果

表 2 主题关联挖掘的结果(部分高关联规则展示)

Lhs(left-hand-side)		Rhs(right-hand-side)	lift
{贸易, 丝绸之路}	=>	{构想}	4.598558
{贸易, 丝绸之路经济带}	=>	{构想}	4.598558
{合作, 贸易}	=>	{构想}	4.598558
{带来, 全球}	=>	{模式}	4.598558
{环境, 全球}	=>	{方式}	4.598558
{贸易, 丝绸之路, 提出}	=>	{构想}	4.598558
{贸易, 丝绸之路经济带, 提出}	=>	{构想}	4.598558
{国家, 贸易, 提出}	=>	{构想}	4.598558
{海上丝绸之路, 贸易, 丝绸之路}	=>	{构想}	4.598558
{合作, 贸易, 丝绸之路}	=>	{构想}	4.598558
{经济, 贸易, 丝绸之路}	=>	{构想}	4.598558
{国家, 贸易, 丝绸之路}	=>	{构想}	4.598558
{带来, 环境, 全球}	=>	{模式}	4.598558
{国家, 海上丝绸之路, 合作, 建设, 经济, 丝绸之路, 丝绸之路经济带, 提出}	=>	{海上丝绸之路}	4.568748
{构想, 国家, 海上丝绸之路, 合作, 建设, 经济, 丝绸之路, 丝绸之路经济带, 提出}	=>	{贸易}	4.568748
{带来, 方式, 环境, 阶段, 模式, 市场, 未来, 影响}	=>	{全球}	4.182074
{构想, 国家, 合作, 建设, 经济, 贸易, 丝绸之路, 丝绸之路经济带, 提出}	=>	{贸易}	4.100122

为了进一步分析这些关联规则所包含的隐含主题知识，以改善度的值作为标准对所获得关联规则进行分类分析，进而发现不同改善度所对应的主题关联规则。在提取主题关联过程中，将表 2 中 Rhs 作为主题知识，通过筛选不同改善度的数值，获得相关的数据，如表 3 所示。

表 3　不同强度关联规则对应的主题知识

改善度(lift)取值	主题知识
lift>9	港口，航线，货物，集装箱，通道，口岸，物流，经济带，地区
8<lift<9	企业，铁路，规划
7<lift<8	压力，增速，风险，增长，下行
6<lift<7	基础设施，基金，区域，大盘，上涨，券商，资金，行情，指数
5<lift<6	改革，产业，生态，基地，提升，打造，板块，创新
4<lift<5	方式，模式，构想，环境，未来，贸易，丝路，重点，丝绸之路，丝绸之路经济带，全球，海上丝绸之路，推进
3<lift<4	丝绸之路经济带，海上丝绸之路，加快

从表 3 可以发现，不同改善度对应的主题知识之间存在一定的差异度，在高改善度值的规则中(lift>8)，主题知识主要描述了“港口”、“经济带”、“铁路”、“物流”等内容；在中等改善度值的规则中(5<lift<8)，主题知识主要描述了“基础设施”、“基金”、“产业”、“创新”等内容；而在较低改善度值的规则中(3<lift<5)，主题知识描述的则为“构想”、“贸易”、“丝绸之路”等内容。从改善度的取值可以看出不同强度的关联规则所对应的主题知识，这也体现了有关“一带一路”新闻报道中不同主题关联的强弱。

为了进一步挖掘不同主题知识所对应的描述信息，将表 2 中 Rhs 作为主题知识，将 Lhs 作为对该主题知识的描述，依据本文方法计算获得有关“一带一路”的相关新闻报道的主题知识的描述，如表 4 所示。

从表 4 可以看出，“一带一路”新闻报道的文本信息中相关主题知识的描述信息。从知识的表达角度来看，前关联 Lhs 是对这些关注重点的语义表述，可以看到这些描述具有明显的语义特征。从这些描述中，可以进一步理解每一个主题知识所对应的知识描述。从表 4 中可以发现有关基金的主题知识则包含基础设施建设以及丝路投资项目等，而创新的主题知识主要是产业和生态项目建设的内容等。从这些词的关联关系中可以更好地理解每个主题所对应的知识描述。

表 4　深度挖掘的结果(部分)

主题知识	知识描述
构想	国家，提出，经济，贸易，合作，丝绸之路，丝绸之路经济带，海上丝绸之路
模式	方式，带来，环境，未来，影响，全球，经济，市场，建设
全球	带来，方式，环境，阶段，市场，未来，影响
交流	举办，主题，活动，合作，发展，国家，建设
基金	基础设施，成立，丝路，投资，项目，国家，建设
创新	产业，生态，基地，提升，打造，城市，重点，加快，项目，建设
丝绸之路经济带	构想，贸易，提出，丝绸之路，海上丝绸之路，合作，经济，发展，中国
贸易	国家，构想，合作，建设，经济，丝绸之路，丝绸之路经济带，海上丝绸之路

表 4 的结果实现了特定领域文本集合潜在知识的发现，关联结果实现了语义维度对文本内容的表示，这些信息的提取有助于信息工作者发现文本隐含的、有价值的知识，也有助于对特定领域知识的深入解读。可见，通过关联规则不仅可实现在海量文本中提取知识，而且能有效地实现知识之间语义的描述。

5　结　语

本文提出将主题模型与关联规则相结合的处理方法，用于挖掘大量文本中所隐含的主题联系，借助主题模型实现文本在语义空间的描述，并成功降维，借助关联规则的方法进一步挖掘文本主题的语义关联。最后，通过实验得到了相关结论。本文提出的方法，将丰富文本信息的知识挖掘思路，并有助于信息工作者更有效地分析大量文本所隐含的知识。

参考文献：

[1] Lazer D, Pentland A, Adamie L, et al. Computational Social Science [J]. Science, 2009, 323(5915): 721-723.

[2] Salton G, Wong A, Yang C. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[3] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval [C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.1998: 275-281.

[4] Agrawal R, Imieliński T, Swami A. Mining Association Rules

Betweensets of Items in Large Databases[C]. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. 1993: 207-216.

[5] 王鉴全，季绍波．基于关联规则的自动构词算法研究[J]. 计算机科学，2014，41(11): 256-259. (Wang Jianquan, Ji Shaobo. Research and Application on Auto-word Building [J]. Computer Science, 2014, 41(11): 256-259.)

[6] 何玉，冯剑琳，王元珍．基于最大关联规则的文本分类[J]. 计算机科学，2006，33(11): 143-145. (He Yu, Feng Jianlin, Wang Yuanzhen. Text Classification Based on Maximal Association Rule [J]. Computer Science, 2006, 33(11): 143-145.)

[7] Cherfi H, Napoli A, Toussaint Y. Towards a Text Mining Methodology Using Association Rule Extraction [J]. Soft Computing, 2006, 10: 431-441.

[8] Sekhavat Y A, Hoeber O. Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views [J]. International Journal of Intelligence Science, 2013, 3(1): 34-49.

[9] 刘菲，黄萱菁，吴立德．利用关联规则挖掘文本主题词的方法[J]. 计算机工程，2008，34(7): 81-83. (Liu Fei, Huang Xuanjing, Wu Lide. Approach for Extracting Thematic Terms Based on Association Rules [J]. Computer Engineering, 2008, 37(4): 81-83.)

[10] Maedche A, Staab S. Discovering Conceptual Relations from Text [C]. In: Proceedings of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Germany. 2000: 321-325.

[11] Schutz A, Buitelaar P. RelExt: A Tool for Relation Extraction from Text in Ontology Extension [C]. In: Proceedings of the 4th International Semantic Web Conference. 2005: 593-606.

[12] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3(3): 993-1022.

[13] Zaki M J. Scalable Algorithm for Association Mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(3): 372-390.

[14] 吴永梁，陈炼．基于改善度计算的有效关联规则[J]. 计算机工程，2003，29(8): 98-100. (Wu Yongliang, Chen Lian. Valid Association Rules Based on Lift-calculation [J]. 2003, 29(8): 98-100.)

作者贡献声明:

阮光册：文献调研与整理，提出研究思路，起草论文；
夏磊：收集数据，实验分析，论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，E-mail: cgruan@infor.ecnu.edu.cn。

[1] 阮光册，夏磊. userdict.txt. 用户自定义词典.

[2] 阮光册，夏磊. model-final.twords. 主题结果.

收稿日期: 2016-09-07
收修改稿日期: 2016-10-18

Mining Document Topics Based on Association Rules

Ruan Guangce　Xia Lei
(Department of Information Management, East China Normal University, Shanghai 200241, China)
(Shanghai Library, Shanghai 200031, China)

Abstract: [**Objective**] This study is to accurately identify potential knowledge correlations among textual information, and then enrich the methodology of knowledge mining. [**Methods**] First, we combined the topic model and association rules. Second, used the LDA model to extract topic set from the texts, which not only reduced the textual dimension but also realized the semantic space expression. Finally, we analyzed the semantic ties among the topics with association rules. [**Results**] We effectively found the potential knowledge association from the document texts with reasonable degrees of support and confidence, and then improved model's "understanding" of the textual message. [**Limitations**] While preprocessing data, the self-defined dictionary posed some negative effects to the results. [**Conclusions**] The proposed method could extract the latent semantic association from unstructured textual information, and then improve the performance of knowledge discovery systems.

Keywords: Association rules　Topic model　Text topics